Strategies for Stabilizing Tomato Prices in Nepal

Project Motivation

The tomato industry plays a crucial role in Nepal's economy, particularly for smallholder farmers who depend on year-round cultivation to support rural employment and meet market demands. Tomatoes are a key part of the Nepali diet, offering essential nutrients and contributing to food security due to their adaptability to various growing conditions.

However, tomato production in Nepal faces significant challenges, including price fluctuations, seasonal changes, and post-harvest losses caused by an inefficient supply chain and unpredictable weather. These challenges affect different types of tomatoes—**Big Tomatoes** and **Small Tomatoes**—in unique ways, as their prices and market dynamics often diverge. For instance, Big Tomatoes tend to have higher prices but are more volatile, while Small Tomatoes are relatively stable but may experience lower demand. Understanding these differences is essential for more accurate price forecasting and targeted interventions.

Improvements in **price forecasting models** can help stabilize incomes for farmers by reducing the uncertainty surrounding price trends for both Big and Small Tomatoes. Additionally, investments in infrastructure and technology can enhance supply chain efficiency, minimize post-harvest losses, and ensure better price stability for both varieties.

Despite these challenges, tomatoes—both Big and Small—hold significant export potential, particularly to neighboring countries. By addressing inefficiencies and improving market systems, Nepal can leverage this valuable crop to promote food security, enhance rural livelihoods, and drive economic stability.

Project Objective

The goal of this project is to create a reliable framework for forecasting tomato prices in Nepal, specifically for Big Tomatoes and Small Tomatoes. By reducing price volatility, improving market predictability, and examining the relationships between tomato varieties, this project seeks to equip farmers, policymakers, and other stakeholders with the insights to make informed decisions and support key players in the Nepalese agricultural market.

Tomato prices in Nepal fluctuate unpredictably due to seasonality, external shocks, and supply chain inefficiencies, impacting farmer incomes, consumer affordability, and overall food market stability. Through accurate price forecasting and volatility analysis, this project aims to mitigate uncertainty, enabling farmers to plan production and harvest schedules effectively. Additionally, by supporting policymakers in developing market stabilization strategies, this study strives to establish a more secure and stable food market in Nepal.

Methodology

To achieve these objectives, this study will employ a structured and multifaceted approach that combines exploratory data analysis (EDA), statistical forecasting models, and volatility analysis.

The project begins with data collection and preparation, using historical tomato price data obtained from local agricultural markets. This data includes average, minimum, and maximum prices, categorized by date, tomato type (Big and Small), and variety (e.g., Nepali, Local). The dataset will be cleaned and organized to ensure it is complete, accurate, and ready for analysis.

The next phase involves Exploratory Data Analysis (EDA) to uncover insights into price dynamics:

- **Historical price trends** will be visualized to identify patterns, cycles, and seasonal fluctuations for Big and Small Tomatoes.
- **Correlation analysis** will examine relationships between price metrics (e.g., minimum, maximum, and average prices) for different tomato types.
- **Principal Component Analysis (PCA)** will reduce dimensionality and identify the most significant factors driving price variations across tomato categories and varieties.

Building on these insights, the study will apply a range of statistical forecasting models to predict future tomato prices, considering both short-term and long-term trends:

- Linear Regression Models: Models will analyze average prices across categories (Big and Small), interactions, and varieties to determine the influence of date, volatility, and categorical factors on price behavior.
- 2. Time Series Models: ARIMA (AutoRegressive Integrated Moving Average) and SARIMA (Seasonal ARIMA) models will capture trends, seasonality, and autocorrelations within the price data.
- 3. **VECM (Vector Error Correction Model)**: VECM will analyze the **cointegrated relationship** between **Big Tomato** and **Small Tomato** prices, identifying their long-run equilibrium and short-term adjustments. Cointegration tests will first confirm this relationship before applying the VECM.
- GARCH (Generalized Autoregressive Conditional Heteroskedasticity) Models: GARCH will capture and forecast price volatility, highlighting clustering effects and short-term risks in price fluctuations.

To ensure robust evaluation, the dataset will be split into **training (80%)** and **testing (20%)** subsets. Forecast accuracy will be evaluated using key error metrics:

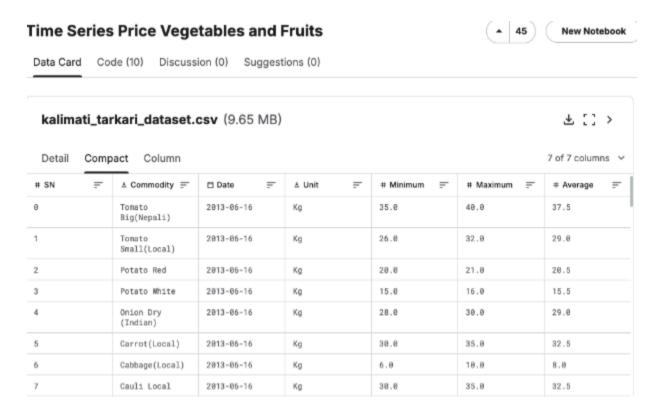
- Root Mean Squared Error (RMSE)
- Mean Absolute Error (MAE)

- Mean Squared Error (MSE)
- Mean Absolute Percentage Error (MAPE)

Forecasts will be compared to actual prices, with results visualized to highlight model performance and reliability. Additionally, **volatility forecasts** from the GARCH model will quantify short-term price risks, supporting stakeholders in managing uncertainties.

Dataset

Kaggle: https://www.kaggle.com/datasets/ramkrijal/agriculture-vegetables-fruits-time-series-prices



This dataset provides a comprehensive overview of vegetable and fruit prices in Nepal from 2013 to 2021. It contains 197,161 entries with daily price information for a wide range of produce, including the minimum, maximum, and average prices recorded.

Sourced from official figures, this dataset offers valuable insights into the price dynamics of essential agricultural commodities in Nepal, making it a useful resource for researchers, policymakers, and anyone interested in agricultural market analysis.

Columns Details:

Commodity - Names of major fruits and vegetables

Date - From 2013 to 2021, daily information

Unit - Per Kg

Minimum - Minimum selling price on that day

Maximum - Maximum selling price on that day

Average - Average selling price on that day

Commodities

The dataset, which contained information on 132 different commodities, was expanded to include two new categorical variables: "category" and "variety." These new variables will aid in arranging and filtering the commodities for future analysis.

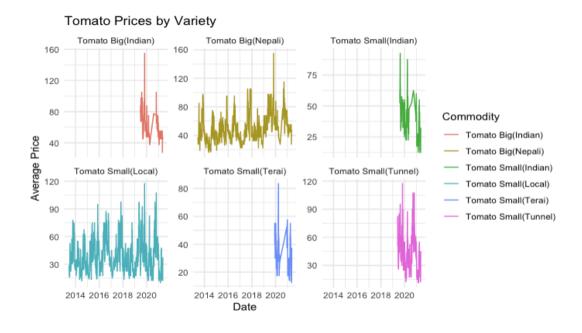
	"Apple(Fuji)"	"Apple(Jholey)"	"Arum"	"Asparagus"
	"Bakula"	"Bamboo Shoot"	"Banana"	"Barela"
[9]	"Bauhania flower"	"Bitter Gourd"	"Bottle Gourd"	"Brd Leaf Mustard"
	"Brinjal Long"	"Brinjal Round"	"Brocauli"	"Cabbage"
[17]	"Cabbage(Local)"	"Cabbage(Terai)"	"Capsicum"	"Carrot(Local)"
[21]	"Carrot(Terai)"	"Cauli Local"	"Cauli Local(Jyapu)"	"Cauli Terai"
[25]	"Celery"	"Chilli Dry"	"Chilli Green"	"Chilli Green(Akbare)"
[29]	"Chilli Green(Bullet)"	"Chilli Green(Machhe)"	"Christophine"	"Clive Dry"
[33]	"Clive Green"	"Coriander Green"	"Cow pea(Long)"	"Cowpea(Short)"
[37]	"Cress Leaf"	"Cucumber(Hybrid)"	"Cucumber(Local)"	"Drumstick"
[41]	"Fennel Leaf"	"Fenugreek Leaf"	"Fish Fresh"	"Fish Fresh(Bachuwa)"
[45]	"Fish Fresh(Chhadi)"	"Fish Fresh(Mungari)"	"Fish Fresh(Rahu)"	"French Bean(Hybrid)"
[49]	"French Bean(Local)"	"French Bean(Rajma)"	"Garlic Dry Chinese"	"Garlic Dry Nepali"
[53]	"Garlic Green"	"Ginger"	"Grapes(Black)"	"Grapes(Green)"
[57]	"Green Peas"	"Guava"	"Gundruk"	"Jack Fruit"
[61]	"Kinnow"	"Kiwi"	"Knolkhol"	"Lemon"
[65]	"Lettuce"	"Lime"	"Litchi(Indian)"	"Litchi(Local)"
[69]	"Maize"	"Mandarin"	"Mango(Calcutte)"	"Mango(Chousa)"
[73]	"Mango(Dushari)"	"Mango(Maldah)"	"Mint"	"Mombin"
[77]	"Mushroom(Button)"	"Mushroom(Kanya)"	"Musk Melon"	"Mustard Leaf"
[81]	"Neuro"	"Okara"	"Onion Dry (Chinese)"	"Onion Dry (Indian)"
[85]	"Onion Green"	"Orange(Indian)"	"Orange(Nepali)"	"Papaya(Indian)"
[89]	"Papaya(Nepali)"	"Parseley"	"Pear(Chinese)"	"Pear(Local)"
Г93 7	"Pineapple"	"Pointed Gourd(Local)"	"Pointed Gourd(Terai)"	"Pomegranate"
[97]		"Potato Red(Indian)"	"Potato Red(Mude)"	"Potato White"
[101]	"Pumpkin"	"Raddish Red"	"Raddish White(Hybrid)"	"Raddish White(Local)"
[105]	"Red Cabbbage"	"Smooth Gourd"	"Snake Gourd"	"Soyabean Green"
[109]	"Spinach Leaf"	"Sponge Gourd"	"Squash(Long)"	"Squash(Round)"
Γ1137	"Strawberry"	"Sugarbeet"	"Sugarcane"	"Sweet Lime"
	"Sweet Orange"	"Sweet Potato"	"Sword Bean"	"Tamarind"
	"Tofu"	"Tomato Big(Indian)"	"Tomato Big(Nepali)"	"Tomato Small(Indian)"
	"Tomato Small(Local)"	"Tomato Small(Terai)"	"Tomato Small(Tunnel)"	"Turnip"
	"Turnip A"	"Water Melon(Dotted)"	"Water Melon(Green)"	"Yam"

	SN <int></int>	Commodity <fctr></fctr>	Date <date></date>	Unit <chr></chr>	Minimum <dbl></dbl>	Maximum <dbl></dbl>	Average <dbl></dbl>	Category <chr></chr>	Variety <chr></chr>
1	0	Tomato Big(Nepali)	2013-06-16	Kg	35	40	37.5	Tomato Big	Nepali
2	1	Tomato Small(Local)	2013-06-16	Kg	26	32	29.0	Tomato Small	Local
3	2	Potato Red	2013-06-16	Kg	20	21	20.5	Potato Red	Other
4	3	Potato White	2013-06-16	Kg	15	16	15.5	Potato White	Other
5	4	Onion Dry (Indian)	2013-06-16	Kg	28	30	29.0	Onion Dry	Indian
6	5	Carrot(Local)	2013-06-16	Kg	30	35	32.5	Other	Local
7	6	Cabbage(Local)	2013-06-16	Kg	6	10	8.0	Other	Local
8	7	Cauli Local	2013-06-16	Kg	30	35	32.5	Other	Local
9	8	Raddish Red	2013-06-16	Kg	35	40	37.5	Other	Other
10	9	Raddish White(Local)	2013-06-16	Kg	25	30	27.5	Other	Local

Exploratory Analysis

Analysis of the dataset revealed that several tomato varieties, particularly Tomato Big (Indian), Tomato Small (Indian), and Tomato Small (Terai), experience substantial price volatility, likely due to factors such as seasonal production, perishability, and localized supply disruptions.

In contrast, Tomato Small (Local) prices are more stable, suggesting consistent supply and demand. Tomato Small (Tunnel) commands higher average prices, potentially due to specialized production techniques like tunnel farming. Both Tomato Big (Indian) and Tomato Small (Tunnel) exhibit short-term price trends.



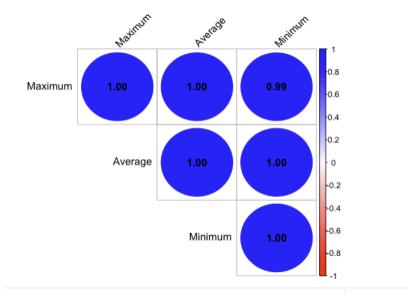
Volatility

The plot reveals that Tomato Big (Nepali) has shown more stable prices since 2016. In contrast, Tomato Small (Local) and Tomato Small (Tunnel) have been more volatile and susceptible to price fluctuations, potentially due to external factors after 2018.

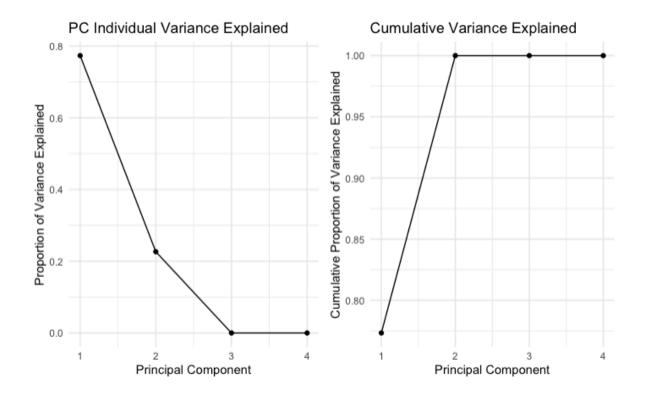


Price Corrplot

There is a strong positive correlation (almost 1.00) that suggests there is a uniform and proportional movement between maximum, average and minimum prices for tomatoes, which implies a stable pricing structure with no major deviation between the different price levels.



Principal Component Analysis



PC1 and PC2 explained 100% of the variance in the Principal Component and PCA Biplot results. PC1 accounted for 77.3% of the variance and PC2 accounted for 22.3%. PC3 and PC4 explained 0 variance and did not meaningfully contribute to the data structure.

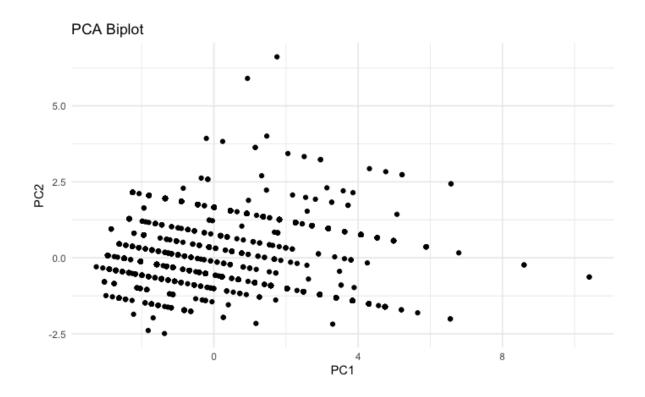
Importance of components:

	PC1	PC2	PC3	PC4
Standard deviation	1.7588	0.9522	2.353e-14	1.739e-15
Proportion of Variance	0.7733	0.2266	0.000e+00	0.000e+00
Cumulative Proportion	0.7733	1.0000	1.000e+00	1.000e+00

Principal Component Loadings:

	PC1	PC2	PC3	PC4
Minimum	0.5596096	-0.18570116	-0.26336058	-0.76354001
Maximum	0.5675875	-0.06165696	-0.54096251	0.61757785
Average	0.5646629	-0.12288251	0.79858291	0.16828847
Volatility	0.2140964	0.97294059	0.01631282	-0.08534189

The first principal component, PC1, is primarily derived from the Minimum, Maximum, and Average values and represents the overall price level. PC2, the second principal component, is derived from the Variability value and represents the price volatility or variability in the tomato price data.



The PCA biplot also shows that data points are spread horizontally along PC1 (price level), and less vertically along PC2 (volatility), indicating that PC1 explains most of the variance. PC2 adds smaller variations. Dimensionality reduction, retaining PC1 and PC2, reduces the dataset from four to two dimensions while preserving 100% of the variance.

Model Selection

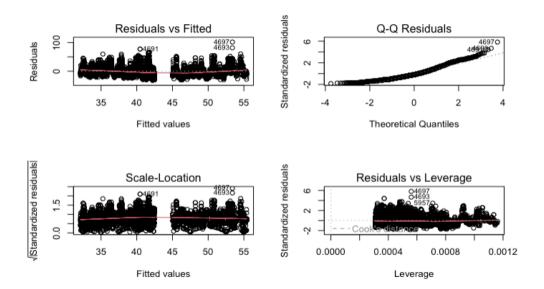
Linear Regression

Using linear regression to forecast tomato prices in Nepal based on the dataset and the created models (category, interaction, variety, and combined) offers several advantages. The linear regression models effectively interpret coefficients, showing the direction and magnitude of the relationships between predictor variables (like category and variety) and tomato prices. Most predictor variables significantly impacted tomato prices (p < 0.05). Multicollinearity was observed in the interaction model, particularly between the Category variable and the interaction term. The Breusch-Pagan test indicated heteroscedasticity (p < 0.05), suggesting that error

variability is not constant across different predictor levels. Lasso regression was used to prevent overfitting and identify the most important predictors of tomato prices. All models were plotted to check for linearity, homoscedasticity, normality of residuals, and influential points.

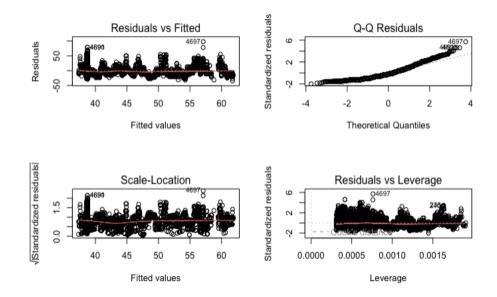
Category Model

Im(Average ~ Date_Num + Category, data = train_data)



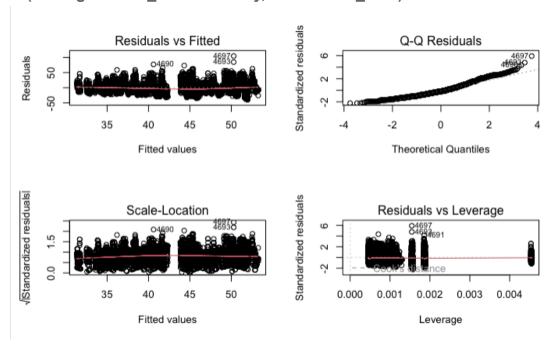
Interaction Model

Im(Average ~ Date_Num * Category, data = train_data)



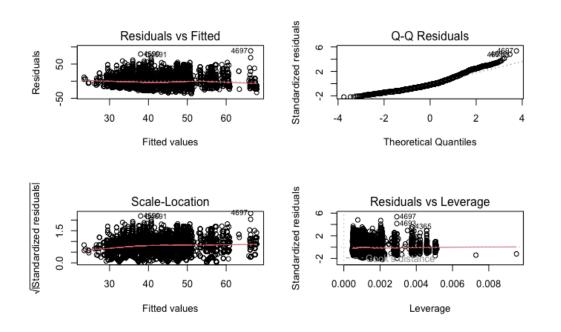
Variety Model

Im(Average ~ Date_Num + Variety, data = train_data)



Combined Model

Im(Average ~ Date_Num + Category + Variety + Volatility, data = train_data)



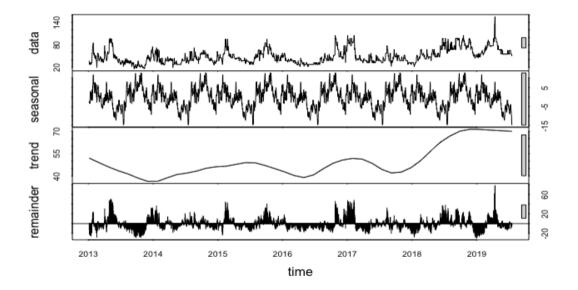
The four models showed varying levels of compliance with linear regression assumptions. The Category model showed the best adherence, with minor issues concerning heteroscedasticity. The Interaction and Combined models showed some improvement in normality but still had issues with non-linearity, heteroscedasticity, and influential data points. The Variety model displayed substantial violations, including non-linearity, non-normality of residuals, heteroscedasticity, and potential outliers. These findings highlight how important it is to critically evaluate model assumptions to ensure valid and reliable results.

Stationarity, Decomposition and Autocorrelation and Partial Autocorrelation

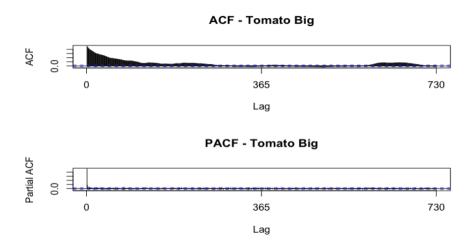
The Augmented Dickey-Fuller Test was used to confirm stationarity; the p-values were below 0.05, indicating that both time series for big and small tomato averages were stationary. Additional analysis, including time series decomposition and examination of ACF and PACF plots, revealed no seasonal patterns and provided potential AR and MA terms for the ARIMA models.

Tomato Big Decomposition and ACF and PCF

This decomposition plot visualizes the different elements that contribute to the price changes of "Tomato Big," and indicates that a SARIMA model should be used, along with other potential factors, to capture the complexities of the data and better predict price fluctuations.

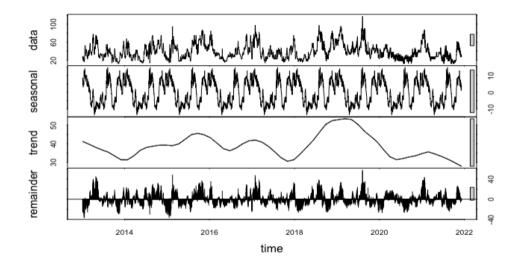


The Autocorrelation Function (ACF) for "Tomato Big" shows significant autocorrelations at multiple lags, particularly at the first few, and with a recurring pattern at lags that are approximately multiples of 365, indicating yearly seasonality. The gradual decay of the ACF suggests potential autoregressive (AR) components. The Partial Autocorrelation Function (PACF) has a significant spike at the first lag and then cuts off sharply, suggesting a potential AR(1) component (one autoregressive term). There are also some significant spikes at lags that appear to be multiples of 365, reinforcing the presence of yearly seasonality.



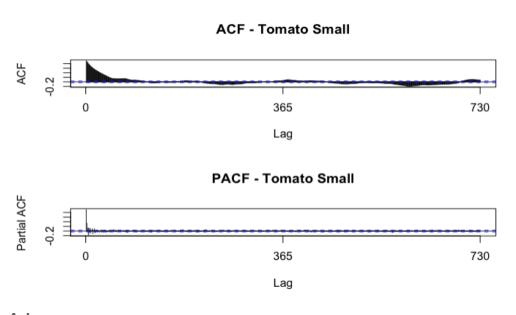
Tomato Small Decomposition and ACF and PCF

The decomposition plot provides a solid foundation for developing an ARIMA model for "Tomato Small" prices. It suggests the potential effectiveness of a SARIMA model, especially with careful attention to seasonal parameters and the incorporation of external factors to address the specific characteristics of this time series.



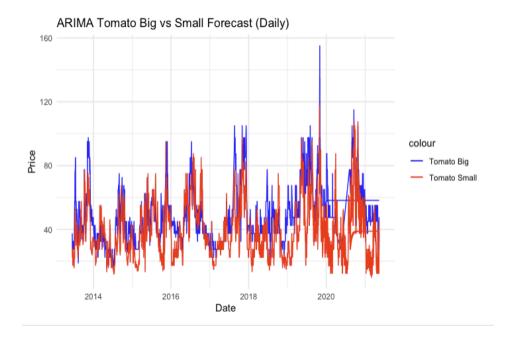
The ACF displays a gradual decay and significant autocorrelations at multiple lags, suggesting potential AR components within the model. Recurring patterns at approximately 365-day multiples point towards yearly seasonality, potentially requiring a seasonal MA component. Additionally, negative autocorrelations at certain lags suggest that a price increase at one point might lead to a price decrease at subsequent points.

Conversely, the PACF displays significant spikes at the initial few lags before cutting off, suggesting potential AR terms, possibly an AR(1) or AR(2) model. Significant spikes at 365-day multiples reinforce the presence of yearly seasonality and the necessity for a seasonal AR component.



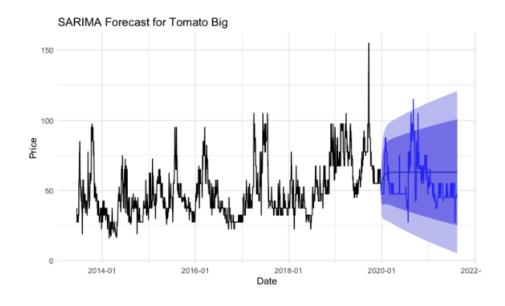
Arima

An ARIMA model was employed to analyze and predict tomato prices in Nepal for "Tomato Big" and "Tomato Small" categories. The analysis revealed that prices for both categories are generally increasing and exhibit seasonal patterns, indicating the cyclical nature of tomato prices, likely influenced by seasonal variations in production and demand. Additionally, "Tomato Big" consistently commands a higher predicted price than "Tomato Small," with this price difference remaining relatively stable over time, despite potential minor fluctuations.

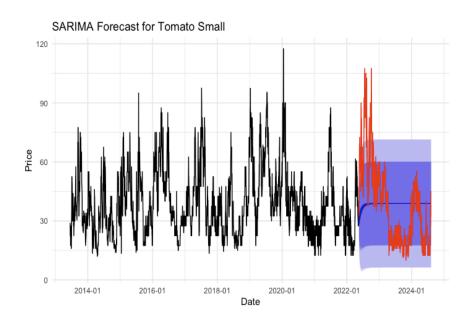


SARIMA

The SARIMA forecast for "Tomato Big" shows a clear upward trend and seasonality prior to 2020. However, the prediction interval is wide and displays significant uncertainty, especially after 2020. The increasing width of the shaded blue area visualizes this uncertainty and indicates decreasing confidence in future predictions as the forecast horizon expands.



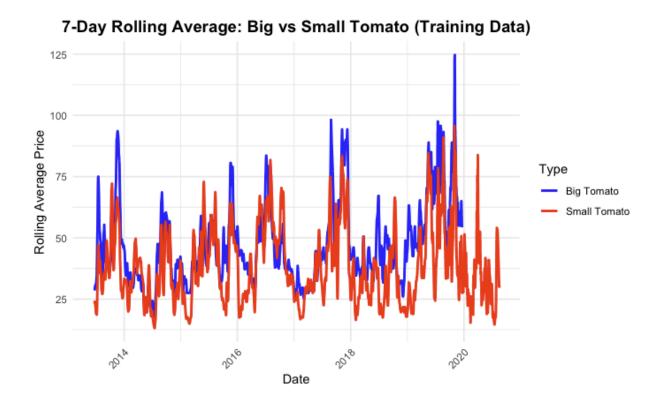
The SARIMA model for "Tomato Small" shows strong seasonality with frequent peaks. To assess the model's performance, actual values (in red) are displayed along with the forecast and compared to observed test data. The prediction intervals, shaded in blue, widen but are slightly narrower than those for "Tomato Big."



The SARIMA model for "Tomato Small" demonstrated a better fit to the data due to consistent seasonality and lower volatility. In contrast, the "Tomato Big" model showed greater uncertainty, likely because of sudden structural shifts or spikes in the data around 2020.

Rolling Average

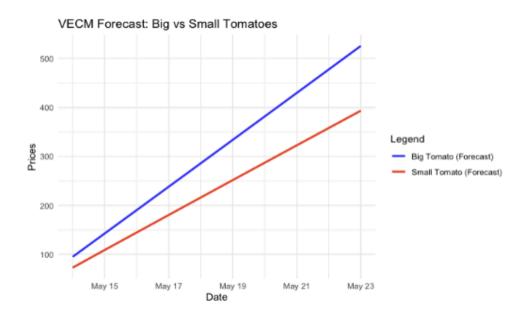
The 7-Day Rolling Average plot reveals clear seasonal patterns for both Big and Small Tomato prices, with regular peaks and valleys over time. While Big Tomato prices are generally higher, sometimes exceeding 100 (especially around 2019-2020), they also exhibit more volatility. In contrast, Small Tomato prices remain more stable and consistently lower, suggesting less susceptibility to market fluctuations. Despite these differences, the overall price trends of both categories appear to be correlated.



VAR and VECM Model and Cointegration

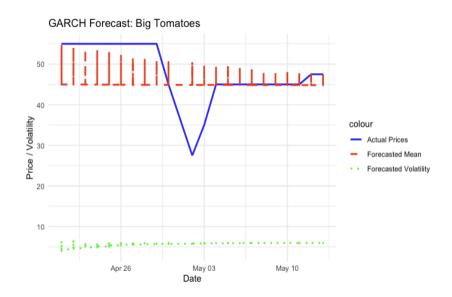
The VECM plot forecasts a short-term, steady upward trend in prices for both Big and Small Tomatoes, with Big Tomato prices consistently exceeding Small Tomato prices. The near-parallel and linear forecasts indicate a stable and proportional relationship between the two tomato types, suggesting that any price deviations are temporary and that a consistent price gap is maintained.

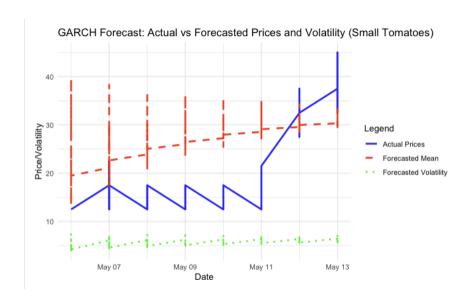
The use of the VECM model highlights evidence of cointegration, indicating a long-run equilibrium relationship between Big and Small Tomato prices. This means that while individual prices may fluctuate in the short term, they will adjust over time to maintain equilibrium. The model captures this long-run dependency while incorporating adjustments for short-term deviations. The linearity observed in the forecast aligns with the short-term nature of the projection, as VECM models tend to perform best in shorter forecast windows when cointegrated series exhibit stable co-movements.



GARCH

GARCH performs better for "Small Tomatoes", where price movements are more gradual. The "Big Tomato" GARCH model fails to capture the abrupt price shock seen in the actual data and assumes stability after the spike. Volatility forecasts remain too low despite the significant price movement, indicating the model may underperform in recognizing extreme variations for Big Tomatoes.





The "Small Tomato" GARCH model identifies a slow upward trend but fails to fully capture the magnitude of sudden price spikes and drops. Volatility forecasts remain consistently low in both cases, indicating a need for more sensitive volatility clustering detection, particularly for datasets with extreme movements.

Results and Findings

The performance of various forecasting models for predicting tomato prices in Nepal was evaluated using RMSE, MAE, MSE, and MAPE metrics. These models encompassed linear regression (including Combined, Category, Interaction, and Variety models), ARIMA, SARIMA, VECM, and GARCH, offering a range of statistical and volatility-based techniques.

Among the linear regression models, the Combined model (incorporating Date, Category, Variety, and Volatility) outperformed the others, exhibiting the lowest RMSE (16.76) and MAPE (35.03%). The Category, Interaction, and Variety models also demonstrated reasonable performance; however, the Combined model's inclusion of multiple explanatory variables enhanced its accuracy.

Model [‡]	RMSE [‡]	MAE	MSE [‡]	MAPE [‡]
Linear Regression Category	17.642938	13.663305	311.2733	38.46027
Linear Regression Interaction	17.379070	13.574792	302.0321	37.62572
Linear Regression Variety	17.920638	13.945077	321.1493	39.70397
Linear Regression Combined	16.767575	12.811861	281.1516	35.03312
ARIMA Tomato Big	14.081899	11.406822	198.2999	20.59743
ARIMA Tomato Small	14.081899	11.406822	198.2999	20.59743
SARIMA Tomato Big	15.426933	13.453333	237.9903	25.79545
SARIMA Tomato Big	15.426933	13.453333	237.9903	25.79545
VECM Tomato Big	10.563711	10.542915	111.5920	22.19561
VECM Tomato Small	10.563711	10.542915	111.5920	22.19561
GARCH Tomato Big	7.348163	4.953835	53.9955	12.24883
GARCH Tomato Small	10.757280	9.487709	115.7192	58.19615

The ARIMA and VECM models effectively captured the price trends and seasonality for both tomato types. The ARIMA models for both Big and Small Tomatoes delivered identical results with RMSE of 14.08 and MAPE of 20.59%, outperforming linear regression models across all metrics. The VECM (Vector Error Correction Model) exhibited significant improvements with an RMSE of 10.56 and MAPE of 22.19% for both Big and Small Tomatoes. The results highlight the importance of capturing the cointegrated relationship between Big and Small Tomato prices, as VECM effectively models both the short-term deviations and long-term equilibrium between the two price series.

Despite being designed for seasonality, the SARIMA model, which extends ARIMA to handle seasonality, did not outperform ARIMA, indicating that seasonal patterns may already be well captured by the ARIMA model. SARIMA showed slightly higher error metrics with an RMSE of 15.42 and MAPE of 25.79%.

The GARCH models showed the best performance in terms of error reduction, particularly for Big Tomatoes. With an RMSE of 7.35, MAE of 4.95, and MAPE of 12.24%, the GARCH model effectively captured volatility clustering and provided highly accurate forecasts. However, for Small Tomatoes, the GARCH model produced higher errors, with an RMSE of 10.75 and MAPE of 58.19%, indicating that price fluctuations for Small Tomatoes are more unpredictable compared to Big Tomatoes.

Conclusion

The results demonstrate that advanced time series models like VECM and GARCH provide superior performance in forecasting tomato prices compared to linear regression and SARIMA models. VECM excels at capturing the long-run equilibrium between Big and Small Tomato prices, achieving consistent error reductions across both categories. Meanwhile, GARCH models effectively handle volatility, particularly for Big Tomatoes, where price variations exhibit strong clustering behavior.

The performance of ARIMA models highlights their robustness in capturing general trends and seasonality in both price series. However, the SARIMA models offered no significant advantage over ARIMA, suggesting limited seasonal influence beyond what ARIMA already captures. The findings suggest that:

- 1. GARCH models are most suitable for price volatility forecasting, particularly for commodities like Big Tomatoes with significant price fluctuations.
- VECM models are ideal for analyzing and forecasting interdependent price dynamics, capturing both short-term adjustments and long-term trends between related price series.

For stakeholders in Nepal's tomato industry, these models provide valuable tools for price prediction and risk management. By leveraging these insights, farmers can plan their production cycles more effectively, and policymakers can address market inefficiencies to stabilize prices and reduce post-harvest losses. This study highlights the importance of combining traditional models with volatility-focused techniques to achieve reliable and actionable price forecasts for agricultural commodities in Nepal.

References

Worldbank: Agriculture, forestry, and fishing, value added (% of GDP) - Nepal

Fairplanet: Innovative AG: How Nepal Tackles Invasive Tomato Pest

ResearchGate: Assessing The Economics Of Production And Marketing Of Tomatoes in Lalitpur, Nepal

Hortjournal: Review on major constraints during off season tomato production in Nepal